One-stop-shop for Artificial Intelligence and Machine Learning for Official Statistics

Project 101146355 – AIML4OS

Workpackage 2

Exploration of AI-ML through citizen science approaches

Deliverable	2.1
Report due	March 2025
Type	R — Document, report
Prepared by	Barry Schouten, Chris Lam
Date prepared	24 March 2025

Workpackage Leader:

Giulio Massacci
giulio.massacci@istat.it

DELIVERABLE 2.1 – CITIZEN-IN-THE-LOOP IN AI/ML

Barry Schouten, Chris Lam (CBS)

SUMMARY: This report is being submitted in accordance with Deliverable 2.1 within workpackage 2 - Communication and Community Engagement. The purpose of this deliverable is to explore the impact that recruiting 'citizens' and/or survey participants (human-in-the-loop) can have on increasing the accuracy of AI/ML models. Based on this assumption, we describe and outline strategies to sample and invite citizens in assisting the creation of AI/ML methods. We distinguish a statistical perspective and an annotator-perspective. Citizen annotators can be involved in three tasks: feature selection, creating training data and updating pre-trained models. Statistical challenges are drift in concepts and features, within-annotator clustering and feature sufficiency. Methodological challenges are annotator bias and annotator drop-out. We discuss these challenges and propose simulations and small-scale qualitative studies within three case studies.

1. INTRODUCTION

Accurate and timely training data are a crucial prerequisite to (supervised) AI/ML methods. For some AI/ML applications training data can be collected at relatively low cost. However, often an investment is needed in evaluating, annotating and checking events of interest. Apart from costs, annotations and corrections may not be feasible without the involvement of the data subjects. This is the setting of this deliverable. We consider official statistics AI/ML (re)training data settings where the help of ‘citizens’ is imperative to obtain accurate training data and to keep training data up to date. As a secondary motivation, especially in the official statistics context, we see added value in involving general populations in terms of transparency, engagement and trustworthiness.

This deliverable is the first of two within the topic of human-in-the-loop. It sets terminology, lines of thought and a framework, and it functions as stepping stone to the second deliverable that includes simulations and small-scale studies. There is a close relation to work done under WP5 of AIML4OS. In particular, the machine learning error framework (Puts, Salgado and Daas 2024) is an important starting point to motivate choices in citizen-in-the-loop strategies.

Citizen-in-the-loop strategies are an interplay between a statistical sampling viewpoint and a methodological annotator recruitment viewpoint. The sampling viewpoint aims at efficacy of citizen involvement, i.e. AI/ML performance. The annotator viewpoint considers annotator psychology in terms of burden and competence. Together they lead to a focus on efficiency: how to reach and maintain sufficient performance while not overloading or overburdening citizens.

Citizen-in-the-loop strategies amount to a special form of active learning. Over the last ten years there has been a strong interest in efficient training and the construction of so-called gold standard corpora in training data, e.g. al-Jarrah et al (2015), Wissler et al (2014), Tyagi and Mittal (2019), Mirzasoleiman, Bilmes and Leskovec (2020) and Monarch (2021). Cacciarelli and Kulahci (2024) give an overview of strategies in active learning. New,

however, is the focus on data subjects as experts. In this context, the communication strategy and citizen engagement become key. Consequently, there is a strong sense of citizen science, e.g. Vohland et al (2021), although the topics do not necessarily have a strong general population interest. Our objectives resemble that of Simson et al (2025). They also consider AI/ML design decisions with the involvement of citizens and having transparency and fairness in mind. Their setting is broader than official statistics and citizens are not necessarily the data subjects themselves.

The main elements in our strategy are a distinction of different citizen involvement phases and a trade-off against in-house annotation by staff. We see four phases: feature selection, baseline training, targeted training and updating. Each requires different citizen involvement. The decision as to which annotation queries may be allocated to citizens is set against the proportion which are to be allocated to in-house staff. A key role in the trade-off is annotator bias, i.e. measurement errors made in the annotations.

The deliverable, and more generally this task within WP2, links to other AIML4OS WP's. The clearest link is to WP5 on standards. It is our objective to ultimately create a standard in citizen involvement, i.e. to merge the fields of survey sampling and data collection designs and AI/ML. We evaluate ideas and tactics at the hand of case studies that are relevant to use case WP's. In selecting case studies, we consulted WP coordinators of these WP's.

This deliverable is organized as follows: In Section 2, we introduce case studies. These will be important in explaining ideas and strategies and will be explored in more detail in the second citizen-in-the-loop deliverable. Next, we move to terminology and framework in Section 3. Subsequently, we discuss the main statistical and methodological challenges in Section 4. In Section 5, we start elaborating strategies and translate them to field studies and simulations. We end with a discussion and a detailed sketch of follow-up activities in Section 6.

2. CASE STUDIES

AI/ML applications may differ in at least three properties: time dynamics, annotator burden and annotator complexity. Time dynamics refer to drift in features and drift in concepts, e.g. Bhattacharjee (2024). In time the definition and meaning of both features and concepts may change. The implications are that a trained model may more often present outdated predictions and need more frequent updating. Annotator burden refers to intensiveness and enjoyment of the annotator task. A task may take relatively more time or may be perceived as cognitively more demanding. Finally, annotator complexity links to the required competence and knowledge of the annotator. Some tasks may by themselves be uncommon to an annotator and require a learning period. However, tasks may also be non-central to an annotator, i.e. require more knowledge or recall than an average person may have. Annotator burden and complexity together may lead to annotator drop-out, annotator bias and inaccurate training data.

We have two criteria for choosing case studies. The first is diversity on the three properties. The second is feasibility to perform simulations and/or qualitative studies for the second part of the citizen-in-the-loop research.

In identifying case studies, we consulted all AIML4OS country leads and coordinators of use case WP's. The case study that found support by multiple countries is the classification of products and services according to COICOP (Classification of Individual Consumption by Purpose). ML models for COICOP classification are used within the contexts of the Household Budget Survey (HBS) and in scanner transaction data for consumer price indices (CPI). Both are mandatory by ESS regulations. Within the ESTAT-project Smart Survey Implementation (SSI), smart options for HBS are evaluated and implemented. A second case study comes from earth observation (EO) data which is a use case in AIML4OS. While WP coordinators indicated they train EO ML models with the help of other sources, we do see value in adding classification of spatial objects. We anticipate that the task has very different annotator properties but also is relatively robust to time change. Spatial objects of interest are garden/green areas, solar panels and different types of crops. As a final case study, we choose classification of travel modes and travel purpose. This case study does not link to AIML4OS use cases. Nevertheless, this case study can be seen as part of the overall process of classifying and coding a variable, regardless of the type of data source used. It is, however, central in time use and passenger mobility surveys. Location tracking data, enriched by points-of-interest data, are to be converted to travel modes and travel purposes. Annotated training data have been collected by CBS and ISTAT in project SSI and by CBS in a large scale research program on a national passenger mobility survey. Available data will allow for simulations. The annotator task is assumed to be relatively burdensome.

Table 2.1 shows the three proposed case studies and their anticipated properties. The product-service case study is affected most by time dynamics. Stores constantly introduce new products, change product names and remove products. Since classification is based on the texts these store present on (e-)receipts, a constant updating is imperative. We anticipate that the characteristics of objects in EO-data are relatively stable. The travel mode and purpose case study is most burdensome as the average persons travels makes two to three travels per day with a number of intermediate stops. The EO case study has a low burden as there are only a few objects to annotate. The product-service case study is the most complex for citizens as they are not familiar with the exact definitions of the statistical categorization (COICOP). Furthermore, the classification contains many 'Other' types of categories combining rarely purchased products and services. This means an intermediate step is needed where citizens find a match through a list presented in common, every-day language that by itself is linked to the formal classification. We anticipate that the context need is highest for travel mode and purpose. There are many locations with little or only weakly informative nearby points-of-interest.

Table 2.1: Properties of the proposed citizen-in-the-loop case studies.

Case study	Time dynamics	Annotator		
		Burden	Complexity	Context need
Product-service COICOP	High	Medium	High	Medium
EO-object identification	Low	Low	Medium	Medium
Travel mode & purpose	Medium	High	Medium	High

An alternative to citizen-in-the-loop is (in-house) staff annotation. Let us term this staff-in-the-loop. The same criteria may be considered as for citizen-in-the-loop except for annotator complexity. This is replaced by the need for context, or, more conceptually, the

need for potentially unobserved but relevant features. In Table 2.1, we add a column representing need for context. Especially, the travel mode and purpose case study have a high need for context.

The case studies will be used to fix thoughts in the following sections. We refer to Lugtig, Roth and Schouten (2022), Burger, Boonstra and Van den Brakel (2024), Klingwort, Gootzen, Remmerswaal and Schouten (2025), Remmerswaal, Lugtig, Schouten and Struminskaya (2025) for background to the case studies.

3. CITIZEN-IN-THE-LOOP PHASES AND TASKS

In this section, we set the stage in terminology and thinking. In essence, our goal is to minimize errors in the machine learning error framework. We define types of training data, human-in-the-loop phases and tasks, and data collection options. We divide human-in-the-loop in citizen-in-the-loop and staff-in-the-loop as discussed in Section 2.

AI/ML training data can be divided into two types. The first is the provision of break/intervention points that demarcate sections of data with the same (anticipated) label. The second is the provision of labels that form the basis to classification. We call the types identification and classification. The case studies vary in the need for both types of data. The travel case study has both: Location data form time series that need to be decomposed into alternating series of tracks and stops. The resulting time sections then are fed into models to predict stop purpose and travel mode. The EO case study also has both but the line between demarcation and categorization is fuzzy. In the product-service case study, the emphasis is on labels and categorization. Splitting up pixels into symbols and words is a relatively established and well-trained field of AI/ML. Nonetheless, errors in text extraction and language processing obviously affect classification.

We see three main phases: The first phase is feature selection, the second phase is training and the third phase is updating or retraining. In the first phase, there are only general ideas about relevant (and available) features. Citizens may be asked how they make decisions, i.e. on what basis they demarcated data sections and on what basis did they annotate these sections. This is particularly relevant to discern features that are being employed by citizens that are not available for an ML model. These may point at unobserved information that leads to noise in training data. Such noise will limit the maximal performance of models, which will be important in monitoring convergence of performance. In the second phase, there is not yet a trained model or optimized set of decision rules but the features have been selected. These features are, in general, not the same as those used by citizens. In the third phase, citizens evaluate predictions of a pre-trained model or set of decision rules and adjust them if needed.

There are a number of important steps that we ignore for now. Features need to be engineered. An ML-modelling approach needs to be chosen. Training data needs to be pre-processed. We assume that these steps follow best practices in AI/ML. More generally, we see a strong resemblance between the process of creating statistics through AI/ML approaches and questionnaire (re-)design, testing and analysis in surveys.

We anticipate that the second phase of training consists of two subphases. One subphase is where a baseline training data set is created without targeted sampling, and one subphase where training is targeted based on some decision rules. This second subphase amounts to active and online learning strategies (see Cacciarelli and Kulahci 2024). We term these baseline training and targeted training.

The three phases imply different tasks. The first phase amounts to qualitative interviews/questions. The second phase leads to data being shown to respondents. The third phase demands for an application, including a user interface, that provides predictions through a pre-trained set of rules or models.

The tasks can be executed within different data collection strategies. For the feature selection phase, the most obvious approach is qualitative, e.g. through user tests conducted within questionnaire/UI-UX labs. The training phase may be part of citizen-science initiatives, e.g. Simson et al (2025), and/or within pilots and early stages of smart survey applications. The updating phase may follow the same approach as the training phase, but now also implemented, repeated smart surveys may be used as vehicle. The updating phase requires an application that provides real-time predictions.

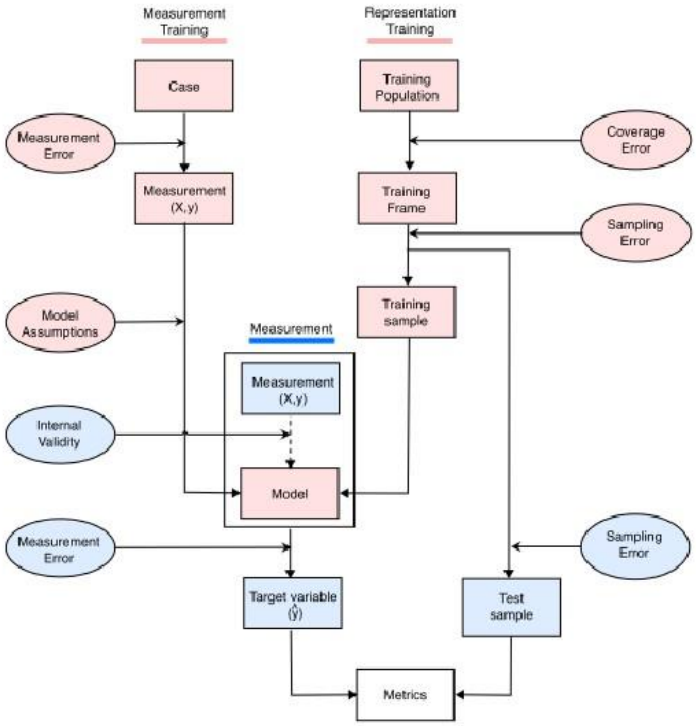
In the training and updating phases the choice is between three options: citizen-in-the-loop, staff-in-the-loop and no action. One of the objectives of our study is to seek decision rules to make these choices.

4. CHALLENGES

Apart from ethical boundaries to large amounts of citizen-in-the-loop requests, there are both statistical and methodological challenges that determine efficiency and efficacy of sampling strategies. We divide them into statistical and methodological challenges as they represent the accuracy perspective and the respondent perspective.

The challenges can be linked directly to the machine learning error framework. See Figure 4.1. We need to address both main ‘arms’, representation and measurement, in the framework. Representation means that no specific citizens should be missing in citizens-in-the-loop either because they are never invited or because they have insufficient motivation or competence. The annotation measurement must be high quality in order to avoid noise and falling performance of predictions, e.g. Sheng, Provost and Ipeirotis (2008).

Figure 4.1: Machine learning error framework taken from Puts, Salgado and Daas (2024). The pink parts concern the training stage and the blue parts the testing stage.



4.1 STATISTICAL CHALLENGES

We see four challenges:

- Friction between citizen features, staff features and data scientist features
- Clustering effects within citizens
- Time shifts in features
- Time shifts in concepts

Friction in features: AI/ML employs features that can be observed/measured for all events to be classified. The ambition is automation, i.e. features must be collected with no or limited additional input from the data subjects. Some features may be collected through introduction/recruitment surveys, but should be stable across the data collection period. An example in the travel case study would be to ask citizens what travel modes (car, e-bike, regular bike, etc) they own and what is their daily activity (study, work, retired, etc). In the product-service case study, the household may be asked at what store(s) they do groceries. Nonetheless, citizens and in-house staff may employ information that is, or even cannot be, captured. This is not necessarily problematic when the features that can be observed have a similar surplus to what a human would use. This will often not be true. There will be a gap in information. This gap is likely even larger for citizens, being the data subjects. The implication is an upper limit to performance of a trained AI/ML. In optimizing convergence, it is crucial side-information.

Clustering: Unlike in-house staff, citizen annotators can only be allocated events for which they are the data subjects. In general, citizens will have a differential importance in the

targeted training and updating phases. Some citizens may be more versatile than others and/or have events that are more outlying in feature distributions. For example, a larger household may be more versatile in products they purchase and outgoing persons may visit more diverse locations. And some persons travel a lot for work or study. There are, however, limits to what can be asked from individual citizens. Clustering effects must, therefore, be accounted for in allocating events to be annotated.

Feature shift: In time, the importance, distribution and even meaning of features may shift. Some feature values may become obsolete while others emerge. For example, products in stores change names, new brands are put on the market and unsuccessful brands may be removed from the (online) shelves. Another example is change of living or working address during the location tracking period. A trained ML model could in such cases even erroneously report high classification probabilities, i.e. not give rise to doubt.

Concept shift: In time, the type of events and categorization may change. New types of events may emerge and others may become rare or obsolete. An example is the rapid growth of e(lectrical)-bikes as a means of transport. E-bikes show similarity to regular bikes in location tracking data, but riders tend to accelerate faster and also travel at higher speed. E-bikes have changed the concept of cycling, regardless of whether they are introduced as a new classification category. Another example could be the introduction of a new type of crop or solar panel in EO-data. As for feature shift, a trained ML model could give erroneous predictions.

4.2 METHODOLOGICAL CHALLENGES

From the respondent-perspective, we see two challenges:

- Annotator drop-out
- Annotator bias/effects

Annotator drop-out: When asking too much from an annotator in terms of amount or frequency of effort, the annotator may drop-out. Drop-out may be selective. Persons with different background characteristics (e.g. age, type of household) may drop-out earlier or later than average, introducing a selection bias. In official statistics setting, representation of the general population is the target. The selective drop-out may be related directly to the events that need to be annotated. A person travelling a lot through different travel modes may get many queries. Persons in households with diverse spending behaviour may get many requests to classify products. Such dependencies imply not-missing-at-random mechanisms in training data.

Annotator bias/effects: Annotators may produce measurement errors. Errors may be random or systematic. Random measurement errors may result from a lack of interest or motivation, leading to so-called ‘satisficing’ behaviour. The resulting noise may lead to a lower accuracy. When annotators differ in the amount of random noise, then they introduce so-called annotator effects. In the decision to invite an annotator, annotator effects may even push more queries to annotators with larger errors. Systematic measurement error implies that annotators consistently choose a wrong classification. This behaviour may be unconscious; the annotator has insufficient knowledge or misinterprets the task. The errors

can also be consciously when the annotator is not willing to reveal certain information. Systematic error may lead to false predictions and reduced accuracy.

4.3 EVALUATION CRITERIA

The training and updating phases amount to optimization problems. Any optimization problem requires explicit objective and cost functions. We discuss options to choose these functions.

The objective function must be linked to the efficacy of annotator sampling strategies. The obvious choice is the performance of the AI/ML predictions based on true/false negatives and positives. Metrics are the F1-score or balanced accuracy. An alternative metric is the absolute calibration bias which is the balanced average of absolute distances between predicted probabilities and observed fractions.

We see multiple cost functions that may be included separately or simultaneously, depending on the context. The main threats to AI/ML performance come from annotator drop-out and annotator bias. Cost functions must, therefore, acknowledge these risks. We have citizen-in-the-loop and staff-in-the-loop. For staff-in-the-loop, annotator drop-out may be assumed negligible but annotator bias must be accounted for. For citizen-in-the-loop both drop-out and bias are realistic. To avoid drop-out, the number of queries, the frequency of queries and the time lags between queries may be bounded. Annotator errors are much harder to translate to cost functions as by their very nature they may go undetected. In part, random errors are caused by lack of motivation similarly to drop-out. Lack of motivation may, thus, to some extent be avoided by constraints on the queries. An option to set some bound to errors that do not result from fatigue, is a limit to variation in annotations for (almost) the same features. A penalty is set on launching queries to feature values that have shown large variations. In this regard, it is crucial to have prior knowledge about the sufficiency of features from the feature selection stage. The variation may also be real when features are insufficient. Another cost function may come from budget. In order to motivate citizens, they may be rewarded with unconditional incentives and/or conditional incentives. For staff, there are inherent costs due to the time they spend on annotating. The budget for citizens and staff will be limited. Given that updating of the AI/ML will often have a longer term time horizon, budget may be specified explicitly per phase.

The annotation sampling strategy is aiming at maximizing accuracy of predictions subject to constraints on number, frequency and time lags of queries, variations in annotations within features and budget for citizens and staff. The optimization of the strategy is discussed in the next Section.

5 PROPOSED STRATEGY FOR ANNOTATOR SAMPLING

5.1 The methodological recruitment perspective

The methodological challenges are in citizen motivation, communication and instruction.

We see five options to recruit citizens:

1. Personal networks: Small-scale recruitment of citizens through the personal networks of NSI staff. Citizens may also include NSI colleagues.
2. Non-probability-based platforms: Invitations may be posted on (semi-)commercial platforms and/or citizen science platforms.
3. Probability-based dedicated sampling: An NSI may draw samples from population registers and send out dedicated invitations to the sampled population units to participate in one or more phases.
4. Follow-up of respondents in repeated surveys: Respondents from relevant repeated probability-based surveys are invited to participate in a dedicated follow-up study.
5. Smart surveys: Annotation is embedded in repeated smart surveys where the corresponding (smart) data are collected by default.

Each option has various design features that need to be further elaborated:

- type and height of incentives;
- form and extent of personal (interviewer) involvement in recruitment;
- form and extent of personal (interviewer) involvement in motivation;
- mode of annotation administration (paper, web-based, mobile app) ;
- pairing of citizens with in-house staff annotators.

Table 5.1: Anticipated fitness -for-purpose per citizen-in-the-loop phase. '+' = well-suited, '+/-' = can be used and '-'= not suited.

	Phase			
Recruitment	Feature selection	Baseline training	Targeted training	Updating
Network	+	-	-	+/-
NPS platform	+/-	+/-	-	+/-
Dedicated survey	-	+	+/-	+/-
Follow-up	+/-	+/-	+	+
Smart survey	-	+	+	+

The fitness-for-purpose of the options varies across the phases. In Table 5.1, we provide the anticipated fitness per phase. Feature selection benefits most from in-depth, qualitative study. This means that relatively a lot of time is asked from a citizen, but that the number of citizens can also be relatively small. The best option is network recruitment and perhaps recruitment through platforms or relevant, repeated surveys. Baseline training is very different and requires both quantity and strong representation. A dedicated survey or a 'smart' survey in which data evaluation and validation by respondents is already implemented seem the best options. Targeted training requires knowledge about past events and context and flexibility to select citizens based on that knowledge. A follow-up survey and a 'smart' diary survey offer such knowledge. A dedicated survey may be an option, but is less flexible as background of citizens is unknown at the invitation stage. For updating trained models all options seem viable. Features may be revisited, large-scale samples may be needed when time dynamics are high, but also flexibility in the selection may be a crucial asset. We conjecture that this phase depends strongest on the characteristics of the case study.

In the second stage of the citizen-in-the-loop project within WP2, the options will be elaborated and evaluated against the three case studies.

5.2 The statistical sampling perspective

Let us suppose that we want to maximize accuracy and that upper limits have been set to cost functions introduced in Section 4.3. To simplify, we assume that budgets are allocated per phase and no overarching optimization is performed. Perhaps at a later stage, optimization strategies may be extended to allow for overall budget constraints rather than budgets per separate phase. Now, how to find an optimal sampling design strategy? To answer this question we consider each of the phases individually. We include ideas from a mathematics challenge conducted January 27-31 by SWI (short for "Studiegroep Wiskunde en de Industrie", Study group Mathematics for the Industry in English). See Dhopeswar et al. (2025) for details. We discuss the trade-off against staff-in-the-loop and the impact of annotator bias in separate subsections.

5.2.1 The feature selection phase

The feature selection phase is qualitative and does not require an explicit sampling strategy. Nonetheless, the selection of citizens must be based on diversity in life style, annotation motivation and annotation competence. These criteria will differ strongly per case study. In the product-service case study a regional spread is important as different store chains tend to operate in different parts of a country. Households from different income quintiles and with different composition should be selected. In the travel case study, a spread across rural and urban areas is key. Persons with different socio-economic status may be selected. In the EO case study again a regional spread is important. Households in different types of dwellings may be selected or farmers with different types of agricultural activity.

5.2.2 The baseline training phase

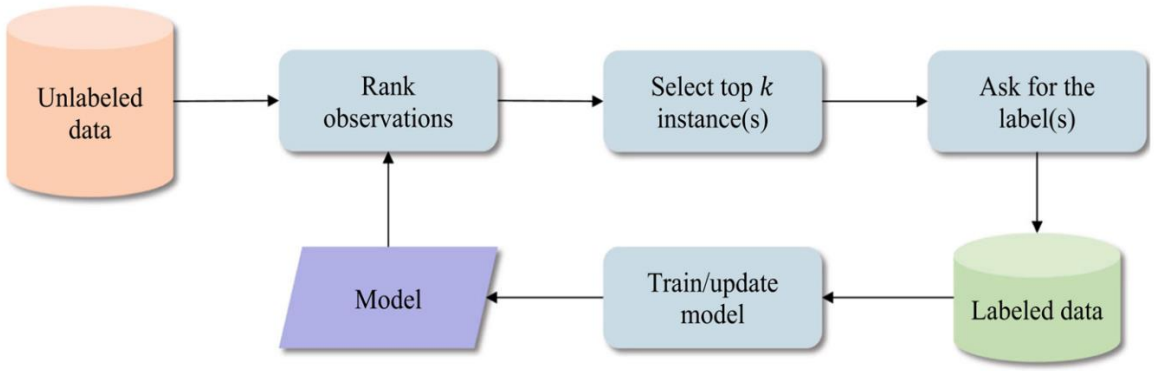
In the baseline training phase there is not yet any targeting based on features. The objective is to create a sufficiently large training data set to get a sense of feature importance and dispersion of the feature space. A general complication is undercoverage of population events in the training data. Essentially, undercoverage amounts to unobserved feature values that correspond to another category set than to the category set they belong too. The existence of such events implies that a ML model needs to be adapted by changing its distance function or by adding more model parameters/levels. However, prior to training data collection, existence of such events may only be conjectured. Hence, in the baseline training phase, it seems to be imperative that sampling is done as randomly as possible. It is the only protection against further strengthening the undercoverage in the successive targeted training phase. Hence, the baseline training phase is about efficiency in terms of certain performance and not about efficacy in terms of optimal performance.

As is true for any form of sampling, setting up an efficient design requires basic prior knowledge of associations. In the case of citizens, it demands for a baseline selection of relevant person/household characteristics. The sampling strategy is, therefore, based on linked background characteristics that substantive experts indicated to be predictive of the

target concepts in the application. Relevant characteristics are very similar to the feature selection phase. The feature selection phase may, therefore, also shed a light on additional relevant background characteristics that had not yet been identified by substantive experts. Demographic and socio-economic characteristics of the citizens may, however, contribute but little to the performance and accuracy. For the stop-track and the product-service case studies it has been observed that person characteristics have relatively low feature importances, for example. This, evidently, leads to a friction as it is the citizens that we sample. There is, for example, a reduced added value in letting a household label all its supermarket receipts for a long time period as it tends to go to the same shop and buy similar types of products. The sampling design strategy must, therefore, acknowledge clustering of features within citizens.

Traditionally, sampling designs are about precision. A sampling strategy is more efficient than another strategy when for a specified required precision the sample size is smaller (or, alternatively, field costs are lower) than for the other. These considerations are the starting point for sample allocations, namely through estimated population variances within population strata formed by the background characteristics. A stratum with a larger population variance would be assigned a larger sample size. Strata with very little to negligible variance would be assigned only a few sample units. In the AI/ML setting, cells would be formed by features. The allocation based on variances amounts to the precision of the predictions. In two strata with the same amounts of training data, the one stratum with prediction probabilities furthest away from 1 will be the least precise. In choosing the allocation of a new sample unit, this stratum should be favoured. More rigorously, the variance may be measured through the prediction probability entropy for a specific subset of units. Since variances are unknown at the onset, a pool/batch-based classification as sketched in Figure 5.1 is a straightforward choice.

Figure 5.1: Algorithm for pool-based classification taken from Cacciarelli and Kulahci (2024).



AI/ML fitting methods have the tendency to overfit when samples are relatively small. This can be adjusted for by applying cross-validation techniques and other types of 'pruning' techniques. We assume that such techniques are successfully applied and that the reported performance and corresponding prediction probabilities do not suffer from a sample-size dependent bias.

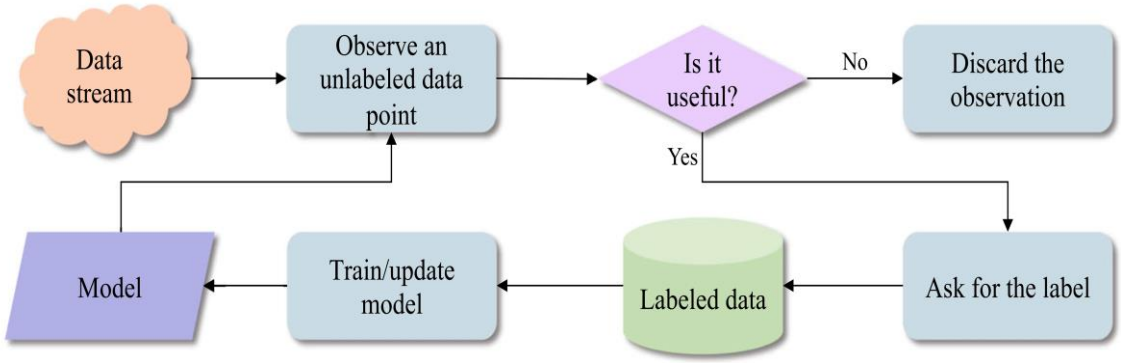
As Figure 5.1 shows, first a pool of data may be collected based on stratified simple random sampling where the strata are defined by relevant characteristics. Annotation is done in batches. When a batch is completed, a new batch may be invited with sampling probabilities based on observed variances. The new batch is included. This process may be repeated until specified variance levels have been reached.

5.2.3 Targeted training phase

In the targeted training phase the focus shifts from efficiency to efficacy. The decision to query an event, i.e. allocate it for annotation, may now be performed on a case-by-case basis based on its added value. The new strategy is sketched in Figure 5.2.

When features are fixed, efficacy points at bias in the potential performance. The training data may suffer from (strong) imbalance and scarcity of certain feature values. The features have not yet been exploited to their full potential. Efficacy may be defined as the change in performance when balancing/weighting training data records. But standard measures such as the F1-score only evaluate performance and give no clues as to how to improve. A measure of utility is needed.

Figure 5.2: Algorithm for instant classification taken from Cacciarelli and Kulahci (2024).



The key lies in feature importance and feature distributions. A growing body of literature is devoted to feature importance through measures such as information gain, Gini gain and SHAP-values. We anticipate that the dimensionality of the feature space and of the classification space are parameters in the speed of convergence of performance. However, these measures, by themselves, are not at the individual event level. Metrics that may be fit for this purpose are based on the distance in features of a new event with respect to annotated events. Dhopeswar et al (2025) consider entropy-based decision rules and uncertainty-based decision rules. They demonstrate that decision rules are far superior to random allocation in terms of the number of events that need to be reviewed.

A complication is that we sample citizens and not events. Given constraints on the number and timing of queries to citizens, it is intractable to decide allocation on an event-basis. A participant may get infrequent and, hence, unpredictable requests. A way out can be to keep track of the utility of events per participant. When the utility is below a specified level for a specified minimum number of events, then the allocation to the participant may be

stopped. Dhopeswar et al. (2025) describe retrospective methods, so-called out-of-bag (OOB) approaches, to determine how important individual participants are in keeping them in the pool of annotators. See Ghorbani and Zou (2019) and Kwon and Zou (2023).

Hence, the procedure sketched in Figure 5.2 may be supplemented by a retrospective step that evaluates the influence of individual annotators.

5.2.4 Updating phase

The updating phase is introduced because of time change, resulting in drift in features and/or drift in concepts. Examples are the inclusion of new stores in the product-service case study, a new type of transport mode in the travel case study and a new type of crop in the EO case study.

Time shifts are a special cases of undercoverage in training data, but with the main difference that undercoverage is now time-dependent. A complicated trade-off emerges between efficiency and efficacy. The reason is that feature importances may change, or, even more rigorously, that new features emerge as being imperative in AI/ML models. Hence, the decision rules as applied during the targeted training phase cannot be applied as is. The metrics may underrate or even overlook events simply because they employed the wrong set of features. For this reason, it seems best to always have a portion of the annotation events that is allocated at random. This portion is needed to revisit the importance of features, and, subsequently, to detect emerging and deteriorating features. In parallel, a portion of the selected events may follow the same targeting strategy as in Figure 5.2, i.e. based on their distance to annotated events. How large both portions should be seems an open research question and one that is likely strongly dependent on the application. We anticipate that in the product-service case study time dynamics are strong and a large portion of sample needs to random. In the other two case studies, time dynamics are likely more modest and more sample may be allocated to targeted updating.

In-time, however, patterns may be distinguished in how feature importances change over time and/or other features are needed. In other words, the feature generation process itself may be modelled. If so, the updating stage may adopt sampling strategies targeting those areas where features are known to change more rapidly.

The clustering of events within citizens may be addressed in analogy to baseline training and targeted training. Hence, in the random updating portion, the focus may be on linked background characteristics. And in the targeted updating portion, it may follow metrics as described in Section 5.2.4. A natural approach to account for the hierarchy of events within data subjects is to introduce a multi-level structure into ML models. To date, this is, however, still a relatively new field of AI/ML.

5.2.5 Staff-in-the-loop

Ideally, the citizen-in-the-loop strategies are merged with staff-in-the-loop strategies. We give a first account of how this could be organized per phase:

- Feature selection phase: Staff annotators may join focus groups and qualitative studies in order to determine whether features they would employ overlap with features that citizens are employing.
- Baseline training phase: The main assumption is that citizen annotators will outperform staff annotators. Part of the sampled events in the baseline training phase may be allocated to staff in order to learn how large the gap is and whether the gap differs across different sections of the feature space. The findings would be the starting point for the subsequent targeted training phase.
- Targeted training phase: The distance of new events can be computed against all annotated events, against events annotated by citizens and against events annotated by staff. The decision to target may then be split into two. Events that are distant to all annotated events are marked for evaluation. If the distance is the similar for staff and for citizen events, then it may be allocated to a staff member. Otherwise the query may be allocated to a citizen participant.
- Updating phase: As for citizen-in-the-loop, it seems best to mix random and targeted sampling.

We note that mixtures of citizen-in-the-loop and staff-in-the-loop are an open area. The case studies in the second stage of the WP2 task will be crucial in finding clues how to proceed.

5.2.6 Annotator error versus time change

So far, we ignored measurement error in annotations. Here, we sketch how detection of such error may be embedded in sampling strategies.

Measurement error and time change are to some extent confounded. Annotated events that appear as outliers in the feature space may be the result of drift in features or concepts. However, in most applications it will be likely that time change is much slower than annotator fatigue. Citizens for which annotations appear as outliers relative to other citizens may be at-risk of error rather than at-risk of being precursors of time change. Nonetheless, it cannot be ruled out that outliers are the first signs of change. For example, the first e-bike owners, the early customers of a new store and farmers experimenting with a new crop would appear as outliers.

As for the targeted training phase, a viable option seems to be to retrospectively look at annotations by individual citizens. Citizens that have a large influence may be selected for review by in-house staff. If annotations are subject to measurement error, then they will deviate from staff annotations.

5.2.7 Summary

Summarizing, we see as tentative strategies to be evaluated in the second stage of the WP2 task:

- Feature selection phase: Qualitative and no explicit sampling strategy. Focus on relevant background characteristics for the concepts of interest, for motivation and for competence.
- Baseline training phase: Random sampling focusing on efficacy

- Targeted training phase: Individual sampling focusing on efficiency
- Updating phase: At first, mix random sampling with targeted sampling similar to training phases, and in time move to stratified sampling.

We have to make one important side remark: There is no guarantee that the amount of required training data has a feasible, manageable size. In the preliminary stage, it may be concluded that the dimensionality/diversity of the 'feature space' and/or the annotation categories are very large. This would then lead to very extensive efforts, involving many citizens. Examples are training of computer vision/pattern recognition models. As a consequence, in the preliminary stage there may be a go – no go decision on whether to proceed to the actual training stage.

6 DISCUSSION AND NEXT STEPS

In this deliverable, we set the stage for pilot studies and simulations. We introduced terminology, four citizen-in-the-loop phases and tentative sampling strategies for each of these phases. The proposed approach resembles active learning but with a focus on data subjects as annotators. The sampling strategies employ linked background characteristics and are a mix of stratified random sampling and targeted sampling based on decision rules that evaluate utility. Extensions to a mix of citizen-in-the-loop and staff-in-the-loop are discussed.

For the second stage of the citizen-in-the-loop WP2 task, we propose:

- Focus groups for case studies product-service and EO to inform the feature selection phase;
- Simulations on case studies product-service and travel to feed the training phase;
- Simulations on case studies EO and travel to feed the updating phase;
- A qualitative study for one of the case studies (yet to be decided which one) as input to the targeted training and updating phases.

Given that the WP2 budget in project days for this task is limited, the focus groups will be organized with the help of NSI colleagues. The qualitative study will, however, employ citizens recruited outside NSI's. The exact format and timing will be decided through a meeting with WP2 coordinators.

REFERENCES

Al-Jarrah et al (2015), Efficient machine learning for big data. A review, Big Data Research, 2, 87 – 93.

Bhattacharjee, S. (2024), Domain Adversarial Learning methods for innovative company classification using website texts, Master Thesis, Technical University Eindhoven, The Netherlands.

Burger, J., Boonstra, H.J., Van den Brakel, J. (2024), Effect of spatial scale, color infrared and sample size on learning poverty from aerial images, Pre-print, Remote Sensing Applications: Society and Environment, Statistics Netherlands, The Netherlands.

Cacciarelli, D., & Kulahci, M. (2024), Active learning for data streams: a survey. Machine Learning, 113 (1), 185-239.

Dhopeswar, R., Dirksen, S., Dubinkina, S., Jurrius, R., Kryven, I., Shen, C., Spitoni, C., Tamang, C. (2025), Reducing annotation burden using active learning, Proceedings of SWI25, January 27-31, 2025, Utrecht, The Netherlands, available at https://www.swi-wiskunde.nl/swi2025.

Elrafey & Wojtusiak (2017), Recent advances in scaling-down sampling methods for machine learning, Wiley Periodicals, 9, Nov-Dec

Ghorbani, A., & Zou, J. (2019), Data shapley: Equitable valuation of data for machine learning, Proceedings of the International Conference on Machine Learning, June 2019, Long Beach Ca, USA.

Juckett, D. (2012), A method for determining the number of documents needed for a gold standard corpus, Journal of biomedical informatics, 45 (3), 460–470.

Klingwort, J., Gootzen, Y., Remmerswaal, D., Schouten, B. (submitted 2024), Algorithms versus survey response: comparing a smart survey travel and mobility app, to appear in Transportation Research Interdisciplinary Perspectives

Kwon, Y., & Zou, J. (2023), Data-OOB: Out-of-bag estimate as a simple and efficient data value, Proceedings of the International Conference on Machine Learning, Honolulu HA, USA.

Lugtig, P., Roth, K., Schouten, B. (2022), Nonresponse analysis in a longitudinal smartphone-based travel study, Survey Research Methods, 16 (1), 17 – 23.

Monarch, R. (2021), Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI, Manning, Shelter Island

Mirzasoleiman, B., Bilmes, J., and Leskovec, J. (2020), Coresets for data-efficient training of machine learning models, Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria.

Puts, Salgado and Daas (2024), Leveraging Machine Learning for Official Statistics: A Statistical Manifesto, paper under review.

Remmerswaal, D., Lugtig, P., Schouten, B., Struminskaya, B. (2025), The effects of study duration on nonresponse and measurement quality in a smartphone app-based travel diary, to appear in Survey Research Methods.

Sheng, V.S., Provost, F., Ipeirotis, P.G. (2008), Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers, Proceedings of KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.

Simson, J., Draxler, F., Mehr, S., Kern, C. (2025), Preventing Harmful Data Practices by using Participatory Input to Navigate the Machine Learning Multiverse, Paper presented at CHI '25, April 26-May 1, 2025, Yokohama, Japan.

Tyagi, S., Mittal, S. (2019), Sampling approaches for imbalanced data classification problems in machine learning, Chapter in Proceedings of ICRIC 2019. Recent Innovations in Computing, p209 – 221, Springer Nature, Switzerland.

Vohland, K., Land-Zandstra, A., Ceccaroni, L., Lemmens, R., Perelló, J., Ponti, M., Samson, R., Wagenknecht, K. (2021), The Science of Citizen Science, Edited Book, Springer Nature Switzerland.

Wissler, L., Almashraee, M., Monett, D., Paschke, A. (2014), The Gold Standard in Corpus Annotation, Proceedings of the 5th IEEE Germany Student Conference, June, Passau, Germany, DOI: 10.13140/2.1.4316.3523